Statistics and Probability Letters 94 (2014) 12–20

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Variable selection in infinite-dimensional problems

Germán Aneiros [a,*], Philippe Vieu [b]

[a] Departamento de Matemáticas, Universidad de A Coruña, Spain
[b] Institut de Mathématiques, Université Paul Sabatier, Toulouse, France

ARTICLE INFO

Article history:
Received 14 November 2013
Received in revised form 24 June 2014
Accepted 28 June 2014
Available online 8 July 2014

Keywords:
Functional data analysis
Variable selection
High-dimensional problem
Partitioning variable selection procedure

ABSTRACT

This paper is on regression models when the explanatory variable is a function. The question is to look for which among the p_n discretized values of the function must be incorporated in the model. The aim of the paper is to show how the continuous structure of the data allows to develop new specific variable selection procedures, which improve the rates of convergence of the estimated parameters and need much less restrictive assumptions on p_n.

1. Introduction

Functional data analysis is a very active field in current statistical researches that presents many different aspects that one could see across the various recent books on the topic (Ramsay and Silverman, 2002, 2005; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012). In this paper one addresses the topic related to regression problems involving some functional covariate χ. In this kind of situation, one has to face with two kinds of questions. As an explanatory purpose one wishes to produce estimates leading to efficient prediction of the response Y, while from an other side as an exploratory goal one wishes to know whether the whole continuous variable χ is influential on Y or whether only a few part of χ can be selected for predicting Y. In a first naive attempt, because the continuous variable χ is always observed through discretized grid, a natural way for modeling the problem would be to think it as a high dimensional variable selection procedure in which the number of scalar explanatory variables is exactly the size of the grid on which χ is observed. In practice the size of the grid (let say p_n) can be much higher than the sample size itself (let say n), and in this sense the problem takes really part among the current active literature on variable selection when having much more variables than individuals (see for instance Candes and Tao (2007), Huang et al. (2008), Meier et al. (2009), Bickel et al. (2009), for a necessarily restricted subsample of this literature).

However, this is not a standard variable selection problem in the sense that the continuous feature of the problem exhibits some strong specificities. The first natural specificity is the possibility of ordering the points in the grid on which χ is observed. Moreover, when the dimension p_n of the problem comes from a finer discretization grid, increasing p_n does not mean adding new information (as in usual high dimensional problems) but means rather having more precise information. The aim of our paper is to show how these specificities can be taken into account in the model to lead to a new kind of variable selection procedure adapted to this functional situation.

The paper is presented in a rather general form allowing for many kinds of estimates and variable selection techniques. It is organized as follows. The model is presented in Section 2 and the so-called partitioning variable selection (PVS) procedure

* Corresponding author.
E-mail addresses: ganeiros@udc.es (G. Aneiros), philippe.vieu@math.univ-toulouse.fr (P. Vieu).

http://dx.doi.org/10.1016/j.spl.2014.06.025
0167-7152/

is introduced in Section 3. This section states our main result (see Theorem 1) in which the asymptotic properties of the PVS method are described. It is worth being noted that this section is written in a general way, without specifying any type of estimate for the parameters of the model. Then, Section 4 will discuss how the PVS method can be applied to many kinds of estimates (penalized least squares and likelihood ones being the most famous) and how the asymptotics obtained in Section 3 exhibit more appealing behaviors (both in terms of rates of convergence and in terms of lower conditions on p_n) than what one would obtain by applying any variable selection procedure not taking into account the functional specificity of the problem. These theoretical advantages are going together with nice finite sample behavior that will be illustrated in Section 5 through some curve dataset analysis coming from chemometrics. This example will show how the PVS method improves upon standard methods which do not integrate the continuity of the data, both in terms of predictive power and in terms of lower computational costs.

2. The model

The statistical sample is composed of n pairs (χ_i, Y_i), $i = 1, \ldots, n$, being i.i.d. as (χ, Y), where χ is a random element of an infinite dimensional space $\mathcal{H}$ and Y is a real random variable. Even if they are intrinsically of continuous nature, the functional variable is only observed through a fine grid. To fix the ideas, in the remaining of the paper one will restrict the purpose to curves data observed on a grid $a \leq t_1 < \cdots < t_{p_n} \leq b$ supposed to be regular in the sense:

$$\exists c_1, c_2, \quad \forall j = 1, \ldots, p_n - 1, 0 < c_1 p_n^{-1} < t_{j+1} - t_j < c_2 p_n^{-1} < \infty. \tag{2.1}$$

Furthermore, the curve χ is assumed to be sufficiently smooth in the sense:

$$\chi \text{ is uniformly continuous on } [a, b], \tag{2.2}$$

and, also, to be bounded away from zero:

$$\exists \eta, \quad \forall t \in [a, b], \ |\chi(t)| \geq \eta > 0. \tag{2.3}$$

In the following, the discretized observations of the curve χ will be denoted by

$$X^j = \chi(t_j), \quad j = 1, \ldots, p_n. \tag{2.4}$$

The regression model can be written as follows:

$$Y = \alpha_0 + \sum_{j=1}^{p_n} \alpha^j X^j + \epsilon. \tag{2.5}$$

In this model it is assumed that only a few part of the curve χ has an effect on the response Y; that is, only a few variables among the p_n ones are really entering into the model (2.5) and this is the meaning of the following assumption:

$$\#S_0 = s_n, \quad \text{where } S_0 = S_{0n} = \{j = 1, \ldots, p_n, \alpha^j \neq 0\}. \tag{2.6}$$

In addition, it is assumed that

$$\exists C, \quad \forall n, \sum_{j \in S_0} |\alpha^j| < C < \infty. \tag{2.7}$$

Looking at the model defined by conditions (2.5) and (2.6) one could think that one has to deal with an usual variable selection problem with high number of covariates. However, because the variables X^j come from the continuous variable χ (see (2.4)), when t_j is close from t_k the two corresponding variables X^j and X^k will roughly contain the same information on the response Y. This is an important specificity of the problem that needs to be incorporated into the model. For that purpose let us introduce two sequences of integers q_n and ω_n ($\omega_n \to \infty$ as $n \to \infty$), and denote by

$$\overline{\alpha}^k_{q_n} = \frac{1}{q_n} \sum_{j=1}^{q_n} \alpha^{j+(k-1)q_n}, \quad k = 1, \ldots, \omega_n,$$

supposing without loss of generality that $p_n = q_n \omega_n$. If we denote:

$$\overline{S}^k = \{j; \ j = 1 + (k-1)q_n, \ldots, kq_n \text{ and } \alpha^j \neq 0\}, \quad k = 1, \ldots, \omega_n,$$

our conditions can be stated as follows:

$$\exists c, \quad \forall j = 1, \ldots, q_n, \ \forall k = 1, \ldots, \omega_n, \ \alpha^{j+(k-1)q_n} \neq 0 \Rightarrow q_n |\overline{\alpha}^k_{q_n}| > c > 0, \tag{2.8}$$

and

$$\forall k = 1, \ldots, \omega_n, \quad \exists 0 < a_k < \infty, \ \overline{\alpha}^k_{q_n} \neq 0 \Rightarrow \#\overline{S}^k \sim a_k q_n \text{ as } n \to \infty. \tag{2.9}$$

The low degree of restriction of these two assumptions will be discussed along Section 4 (see Remark 1). Once q_n and ω_n have been chosen, for any $j = 1, \ldots, p_n$ one denotes k_j the unique integer $k \in \{1, \ldots, \omega_n\}$ such that $j \in \{(k-1)q_n + 1, \ldots, kq_n\}$.